Session 4

ADVANCES IN DATA EDITING

IMPROVING OUTLIER DETECTION IN TWO ESTABLISHMENT SURVEYS

Julia L. Bienias,[1] David M. Lassman, Scott A. Scheleur, and Howard Hogan
U. S. Bureau of the Census

I. Introduction

One step in producing estimates from survey data is editing. In many settings, trained analysts examine the data to find unusual or unexpected values, which may be the result of errors made by the respondent or in the data-capture processes. Having found a questionable case, the analyst then tries to verify its accuracy by checking the original form, obtaining related data from other sources, and/or contacting the respondent. One would like to correct as many errors as possible within the time limitations for a given survey. Thus, accurately identifying the cases whose values are most likely to be the result of errors is an essential part of efficient editing.

Previous researchers have successfully used various graphical methods to improve both the efficiency and accuracy of the editing process (e.g., Esposito, Fox, Lin, & Tidemann, in press; Granquist, 1990; Houston & Bruce, 1992; Hughes, McDermid, & Linacre, 1990). We describe the application of graphical methods from exploratory data analysis to the task of identifying potentially incorrect data points. Our report is the result of a working group of analysts, research statisticians, and programmers devoted to this effort.[2] We illustrate the methods with data primarily from the Annual Survey of Communication Services and the Monthly Wholesale Trade Survey. We first describe the two surveys and the current methods used for editing.

2. Descriptions of the Two Surveys

2.1 The Annual Survey of Communication Services

The Annual Survey of Communication Services (ASCS) is a mail survey covering all employer firms that are primarily engaged in providing point-to-point

[1] This paper reports general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau. Address correspondence to: Julia L. Bienias, Economic Statistical Methods and Programming Division, Bureau of the Census, FOB 3015-4, Washington, DC 20233.

[2] We thank the other members of our working group, without whom the work described here would not have been possible: Thomas Bell, Willard Caldwell, Vicki Garrett, Imelda Hall, Donald Hundertmark, Jennifer Juzwiak, William Knowlton, David Stachurski, and Georgeann Wright. We also acknowledge the many other members of Business Division and of Economic Statistical Methods and Programming Division who have supported this effort and who continue to support it.

communication services (e.g., telephone, television, radio), as defined in Major Group 48 of the 1987 edition of the *Standard Industrial Classification Manual*. The ASCS provides detailed revenue and expense statistics from a sample of approximately 2,000. The Census Bureau introduced the survey in 1991 to track the explosive growth and change in the industry. The Bureau of Economic Analysis is the primary federal user of the data collected; other users are the Bureau of Labor Statistics and private industry (U.S. Bureau of the Census, 1992.)

2.2 The Monthly Wholesale Trade Survey

The scope of the Monthly Wholesale Trade Survey (MWTS) is all employer firms engaged in wholesale trade, as defined by Major Groups 50 and 51 of the 1987 edition of the *Standard Industrial Classification Manual*. Particularly, the survey covers merchant wholesalers who take title to the goods they buy and sell, collecting sales and inventory information. The MWTS, conducted since the 1940's, is a mail survey of approximately 7,000 firms, of which 3,500 receive forms in a given month. As with the ASCS, the Bureau of Economic Analysis is the primary federal user of the data. (See U.S. Bureau of the Census, 1994.)

3. Issues Involved in the Current Editing Procedures

After the data from the questionnaires are keyed, a computer program flags cases failing completeness, internal consistency, and/or tolerance edits. Editing review is divided among several analysts for a given survey. Each analyst finds which edits have failed for a case through an interactive correction system or a paper listing, on a case-by-case basis. They can also use a database query system to try to find problem cases that have not already been identified.

There are several disadvantages to this approach. Examining one case at a time does not permit the analyst to obtain a broad view of the behavior of the industry as a whole, and such a view can be of great benefit in determining the impact of an individual unit on the aggregate estimate. In addition, it undoubtedly leads analysts to examine more cases than necessary. Finally, for a few of the ASCS tolerance edits, constant parameter levels derived from previous surveys have been hard-coded into the programs. This implicitly assumes the relationships among the variables are static over time, which may not be the case.

4. Application of Exploratory Data Analysis Methods

4.1 Background

Exploratory data analysis (EDA) can be described as "a set of tools for finding what we might have otherwise missed" in a set of data (see Tukey, 1977). These tools, combined with the analysts' subject-matter expertise, are particularly well-suited to the task of data editing. In this setting, we are not interested in ascertaining the truth of a postulated economic model or a similar estimation or hypothesis testing problem. Rather, our goal is to determine which cases are unusual with respect to the bulk of the cases and to follow up those cases. In addition to providing methods for displaying data in a variety of ways, EDA emphasizes fitting data using methods that are relatively insensitive to the presence of outliers in the data ("resistant" methods). Such fitting is a way to define and then account for (remove) certain aspects of the data so the analyst can concentrate on other aspects. (See Hoaglin, Mosteller, & Tukey, 1983; Velleman & Hoaglin, 1981.)

EDA fits well with the survey processing environment. Because in the editing stage we expect to find wild observations that might be off by orders of magnitude from the bulk of the data, transformations and resistant techniques are particularly useful in helping us find order amid the chaos. In addition, these techniques allow for efficient examination of large amounts of information at once, an aspect that is particularly valuable in the time- and resource-constrained survey production environment.

From the arsenal of tools collectively called "exploratory data analysis," we considered both univariate boxplots and the more general bivariate fitting. We describe boxplots first, followed by scatter plots and some methods for fitting. In addition, although transformations are applicable to all tools, we describe them in the context of scatter plots, because that is where we used them most.

4.2 Boxplots

Boxplots allow quick visual analysis of the location, spread, and shape of a distribution. Our boxplot has its box spanning the lower and upper quartiles, with whiskers extending from the box to the furthest data point within a distance of one-and-one-half times the interquartile range from the box. We considered data values beyond the whiskers as potential outliers. If the data are reasonably symmetric, then these cutoffs provide a good working definition of cases which may need review. See Tukey (1977) for a discussion of boxplots in general, and Hoaglin, Mosteller, and Tukey (1983) for a discussion of the expected number of outliers for different sample sizes. Note that the whisker definition could be modified to suit the needs of a particular survey operation (e.g., one could use 2 times the interquartile range instead of 1.5).

Figure 1 demonstrates the use of the boxplot for operating ratio (expenses/revenue) data from the ASCS.[3] The boxplot shows that the median operating ratio is .7978 and fifty percent of the points lie between .7269 and .9811. The left and right whisker values are .3760 and 1.3401. The cases flagged by the

[3] To protect the confidentiality of our data, we have not provided details about the particular subset of data analyzed in each plot, nor have we labeled axes when such information could be revealing.

use of the boxplot are different (and fewer in number) than the cases that would have been flagged by the current hard-coded edit parameters, .9 and 1.1. Those parameters fail to help us isolate the "true" outlier cases, as they result in too many cases being flagged. Alternatively, we could flag cases that would appear beyond the whiskers as in our boxplot, an approach that is "dynamic" in that it relies on incoming data to set parameters. At minimum, we could use values from Figure 1 as new hard-coded edit bounds, noting that these revised bounds would no longer be symmetric around one (consistent with the findings of Granquist, 1990).

4.3 Scatter Plots

A scatter plot of two variables is a simple and particularly useful technique. When the data are appropriately transformed, one can use a variety of methods to remove linearity in the scatter and then examine the residuals from the linear fit. This allows us to see patterns that we might otherwise miss when looking at the original data; looking at the residuals from a fit allows us to examine the data on a finer scale (see Section 4.5).

As a vivid illustration of the kinds of problems encountered in editing data, we used another survey for which we had raw responses to a particularly problematic question. One item in the Motor Freight Transportation and Warehousing Survey is the percent of revenue derived from local trucking, a question believed to be confusing to respondents may define "local" in different ways. Figure 2, a scatter plot of these unedited data for the current versus prior period, shows a weak linear relationship. Cases along the 45 degree line are companies whose year-to-year reports are consistent. The reports become more inconsistent the further they are from the 45 degree line. Some of the cases along the vertical axis are "births" to the survey (cases selected during the current period to reflect new firms). Births should be analyzed separately, because they have only current-year data.

4.4 Transformations

Transforming the data so patterns can be more easily discerned is a technique that is important to all graphical and data-fitting methods. It is used to obtain symmetry in the data, to promote linearity, and to equalize spreads between data sets. These properties are assumed, implicitly or explicitly, by many of the techniques we use to analyze data. For example, when we look for outliers by examining a boxplot, we are implicitly assuming the data are supposed to be symmetric. If the data are naturally skewed, many of the points in the tail that appear to be outliers are actually values that are consistent with the underlying distribution. Thus, "discovering" such outliers in the long tail would not be very meaningful. With skewed data, we want to spend our time investigating those data points that are particularly unusual, given that we expect many points far from the bulk of the data. If we transform skewed data to be generally symmetric, we can then find those points.

Because economic data are typically positively-skewed, transformations that lead to the expansion of lower data values and to shrinking the spread of larger data values are particularly useful. (See Hoaglin, Mosteller, & Tukey, 1983, for more details on types of transformations.)

Figure 3 is a an example of the use of transformations for the ASCS. The scatter plot of untransformed revenue data (Figure 3a) reveals little, as one case is many times larger than the other cases. Hiding the large case was unsuccessful, as the next largest case was still many times larger than the remaining cases. Instead, taking logs of the data showed a useful scatter plot (Figure 3b). We see a strong linear relationship, which is what we expect for a plot of current and prior data. Cases that do not appear to be following this linear relationship would thus be considered unusual. We also see that the case that appeared to be an outlier in Figure 3a is, in fact, very much in line with the rest of the data.

For the MWTS, a scatter plot of the current inventory data against the current sales data shows that most of the data are bunched in the lower left corner (see Figure 4a). Because both variables are skewed, we first tried a natural log transformation (log(x + 1)). (We added one because a value of 0 for inventory data does not indicate the case is a birth, and thus it may be of value to include such cases.) This overtransformed the data, skewing them in the opposite direction (Figure 4b).[4] This is because there was a big gap in values between 0 and the next largest value. Such an effect would also occur if there were many establishments with very small reported data and a few with very large values. We then tried taking the square root (Figure 4c) and fourth root (Figure 4d). The latter resulted in the most useful transformation, as most of the data can be seen clearly.

4.5 Fitting

In this section we describe two approaches to fitting, ordinary linear regression and resistant regression. Both were useful, in different ways.

In analyzing ASCS data, we considered the relationship between revenue and payroll for current year data. Figure 5a shows the ordinary least squares regression of revenue on payroll; there are many points clustered near the origin and two cases in the upper right corner. First, we tried removing the two large cases. Again the distribution showed points clustered in the left corner. Such an approach, of iteratively hiding points and refitting, has the disadvantage of being subjective and of essentially requiring analysts to identify outliers first.

One alternative is to use ordinary least squares on transformed data. In this example, logs were useful. Figure 5b shows the fit to the logged data, depicting a

[4] If the cases with 0 reported inventory are ignored, as they might be for other variables, then the logarithm transformation provided a useable picture of the data.

strong linear relationship. The point labeled A is an obvious outlier. Examination of the residuals revealed a pattern, which allowed us to discover that tax-exempt cases were inadvertently being included in the analysis. Tax-exempt cases should be examined separately from taxable cases, because our revenue item only includes taxable receipts. Removing both those cases and point A and refitting the data (Figure 5c) led to the distribution of absolute residuals shown in Figure 5d. This plot can be used to detect outliers, as with a cutoff level C:

$$C = K * \text{(median absolute residual)}.$$

We found $K=4$ (corresponding to $C = .7868$) to be the best. All cases above .7868 were examined and most were "true" outliers. For our example, this method was judged by the survey analysts to be excellent for finding outliers.

Unfortunately, ordinary least squares (OLS) can give great weight to fitting a few wild values. It may work well, as in our example, when there are only a few wild cases and the demarcation between usual and unusual is clear. As an alternative, we investigated resistant fitting using the biweight function developed by Tukey (Mosteller & Tukey 1977; McNeil, 1977). This widely-tested iterative weighted-least-squares fitting procedure uses a weighting function defined as:

$$W_i = \begin{cases} (1-u_i^2)^2, & u_i<1 \\ 0 \text{ otherwise,} & \end{cases}$$

where $u_i = (r_i / (c*s))$

$r_i \equiv$ Residual from previous fit for point i

$s \equiv$ mean absolute residual from previous fit

$c \equiv$ scale factor.

Setting $c = 4$ is quite resistant, $c = 8$ is moderately resistant. We stopped iterating when the proportionate change in s was less than 0.01. This required few iterations; resistant regression is a very efficient and fast procedure.

We applied resistant regression to the MWTS, predicting logged current inventory data by logged inventory data from the prior year. We expect a linear relationship. Figure 6a shows the data and the line from the OLS fit, and Figure 6b shows the residuals from that fit. It is easy to see the OLS fit misses the central tendency of the point cloud. Figure 7a shows the fit resulting from resistant regression ($c=4$). This fit more effectively removed the linearity from the data. The residuals now cluster around 0, as we would want (Figure 7b).

5. A Note on Using Ratios

In many instances, data review has relied on calculating ratios (e.g., sales/payroll) and looking for unusually large or small ratios. There is nothing wrong with this approach per se, but it would be wrong to rely too strongly on it.

The use of ratios assumes a rather simple model of the true relation between the two variables, specifically a straight line through the origin. The true relation may differ markedly, there may be data clouds following different straight lines. For example, the relationship might be different for a small company than for a large company. It is essential that the data reviewer plot the data and look at the shape. Further, the "acceptable ratios" are often set from historic data, last year's or last census'. The relationships can change systematically throughout the business cycle. One could iterate, calculate the average ratio from the current survey, calculate its standard deviation, identify and remove outliers, and start again. However, given the existence of rather fine iterative resistant fitting tools, it is hard to see the advantage of this approach.

6. Summary and Extensions

We have described how principles and methods from EDA can be used to improve the efficiency and accuracy of editing, by helping analysts see patterns in the data and use that information to prioritize cases for follow-up. Building a successful editing system using this approach is more than just selecting the correct statistical tools. The system must be acceptable to the people who will use it. Creating such acceptance requires training the analysts in the methods described here, as well as incorporating the tools into the current production environment and existing computer systems. To date, we have been successful in getting many people to try the methods on several surveys. In addition to the surveys described previously, these methods are currently being applied to the Motor Freight Transportation and Warehousing Survey, the Service Annual Survey, and the Commodity Flow Survey.

Analysts for these surveys reported that being able to ascertain the effect of a given case on the estimate was quite useful. Other specialized programs written for data editing provide this feature (e.g., Esposito, Fox, Lin, & Tidemann, in press; Houston & Bruce, 1992). Incorporating sampling weights in the procedures described here provides a similar utility.

The EDA approach can be combined with batch-type edits (e.g., SPEER, Draper, Greenberg, & Petkunas, 1990; Lee, in press). One could examine the data flagged from a batch program along with the unflagged data using the tools described here. Or, the graphical-based methods could be the basis for batch-type dynamic edits. For example, a program could transform the data to be more symmetric and then flag all cases that would be beyond the whiskers of a boxplot. Finally, in settings in which hard-coded edit parameters must be used, these methods can be used on a subset of data to help find or evaluate such cutoffs.

7. References

Draper, L ., Greenberg, B., & Petkunas, T. (1990). On-line capabilities in SPEER (Structured Programs for Economic Editing and Referrals). *Proceedings of Statistics Canada Symposium 90: Measurement and Improvement of Data Quality*, pp. 235-44. Ottawa: Statistics Canada.

Esposito, R., Fox, J. K., Lin, D., & Tidemann, K. (in press). ARIES: A visual path in the investigation of statistical data. *Computational and Graphical Statistics.*

Granquist, L. (1990). A review of some macro-editing methods for rationalizing the editing process. *Proceedings of Statistics Canada Symposium 90, Measurement and Improvement of Data Quality*, pp. 225-34. Ottawa: Statistics Canada.

Hoaglin, D.C., Mosteller, F., & Tukey, J. W. (Eds.) (1983). *Understanding Robust and Exploratory Data Analysis*. NY: Wiley.

Houston, G., & Bruce, A. G. (1992, February). Graphical editing for business and economic surveys. Technical report, New Zealand Department of Statistics, Mathematical Statistical Branch.

Hughes, P.J., McDermid, I., & Linacre, S. J. (1990). The use of graphical methods in editing (with discussion). *Proceedings of the 1990 Bureau of the Census Annual Research Conference*, pp. 538-54. Washington, DC: U.S. Department of Commerce.

Lee, H. (in press). Outliers in survey sampling. In B. Cox et al. (Eds.), *Survey Methods for Business, Farms, and Institutions*. NY: Wiley.

Mosteller, F., & Tukey, J. (1977). *Data Analysis and Regression*. Reading, MA: Addison Wesley.

McNeil, D. R. (1977). *Interactive Data Analysis*. NY: Wiley.

Office of Management and Budget. (1987). *Standard Industrial Classification Manual.* Available from National Technical Information Service, Springfield, VA (Order no. PB 87-100012).

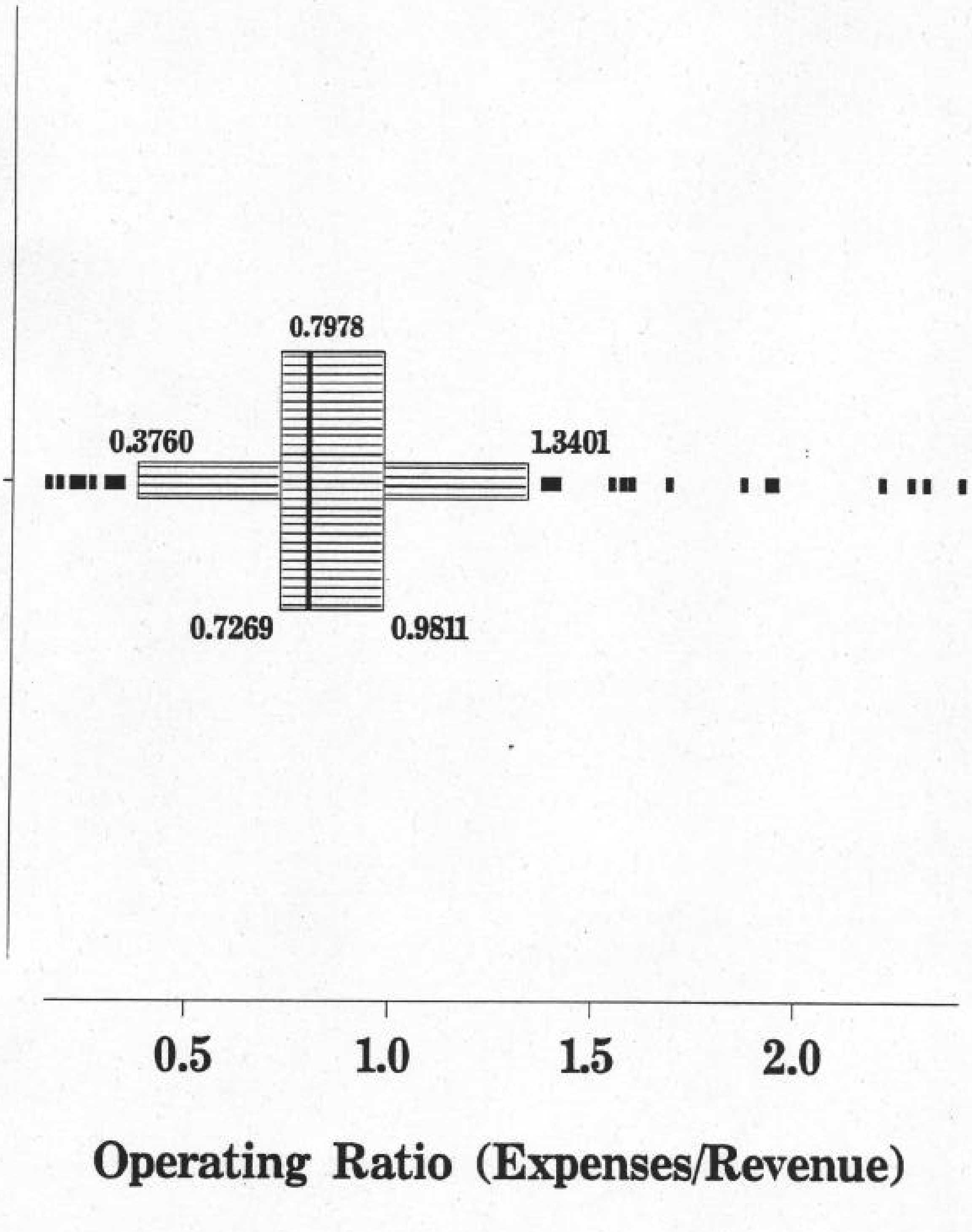
Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

U.S. Bureau of the Census. (1992). *Annual Survey of Communication Services: 1992* . Washington, DC: U.S. Government Printing Office (Current Business Reports, Item BC/92).

U.S. Bureau of the Census. (1994, April). *Combined Annual and Revised Monthly Wholesale Trade, January 1987-December 1993.* Washington, DC: U.S. Government Printing Office (Current Business Reports, Item BW/93-RV).

Velleman, P.F., & Hoaglin, D. (1981). *Applications, Basics, and Computing of Exploratory Data Analysis*. Boston: Duxbury Press.

Figure 1


0.7978
0.3760
1.3401
0.7269
0.9811
0.5
1.0
1.5
2.0
Operating Ratio (Expenses/Revenue)

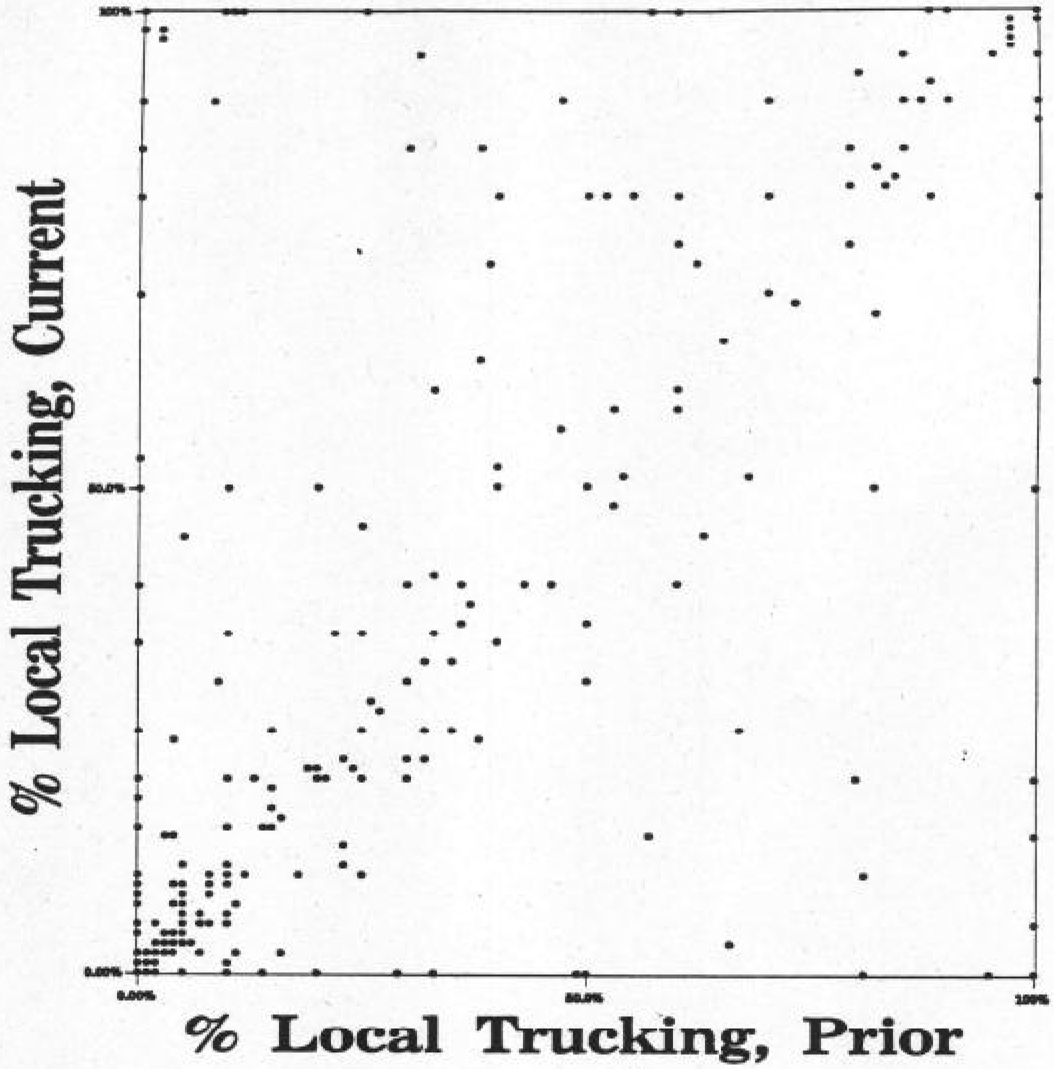

100%
50.0%
0.00%
0.00%
50.0%
100%
% Local Trucking, Current
% Local Trucking, Prior

Figure 2

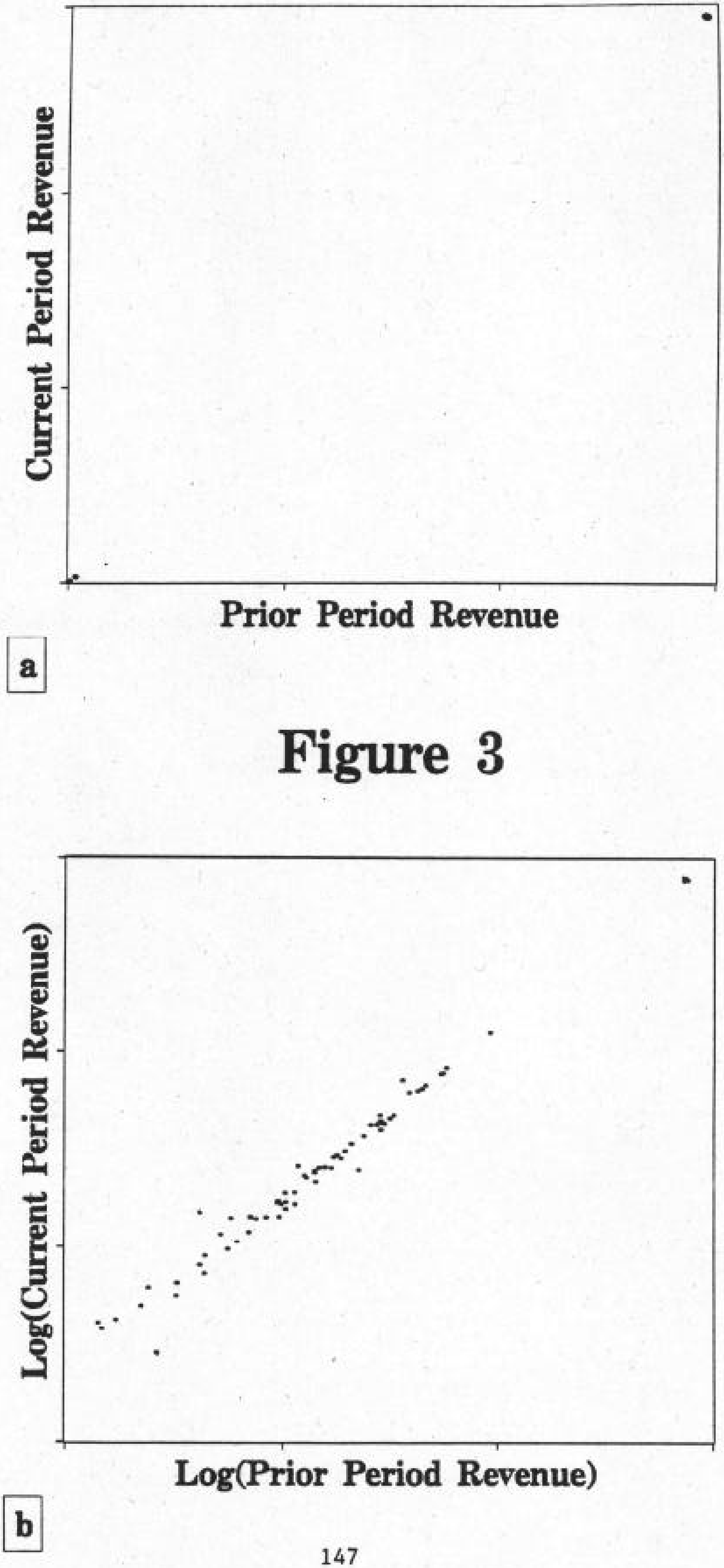

Current Period Revenue
Prior Period Revenue
a
Log(Current Period Revenue)
Log(Prior Period Revenue)
b

Figure 3

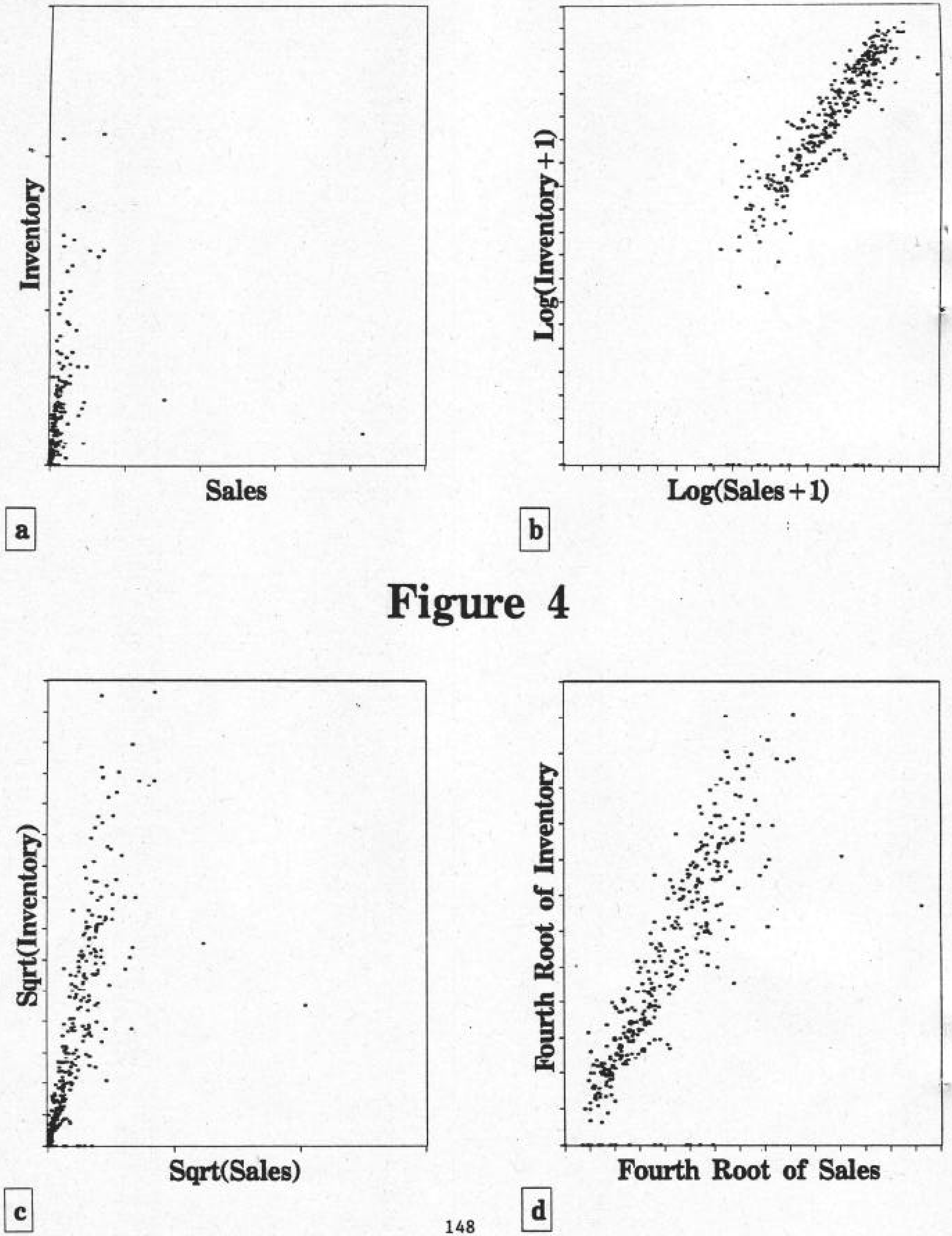

Inventory
Sales
a
Log(Inventory+1)
Log(Sales+1)
b
Sqrt(Inventory)
Sqrt(Sales)
c
Fourth Root of Inventory
Fourth Root of Sales
d

Figure 4

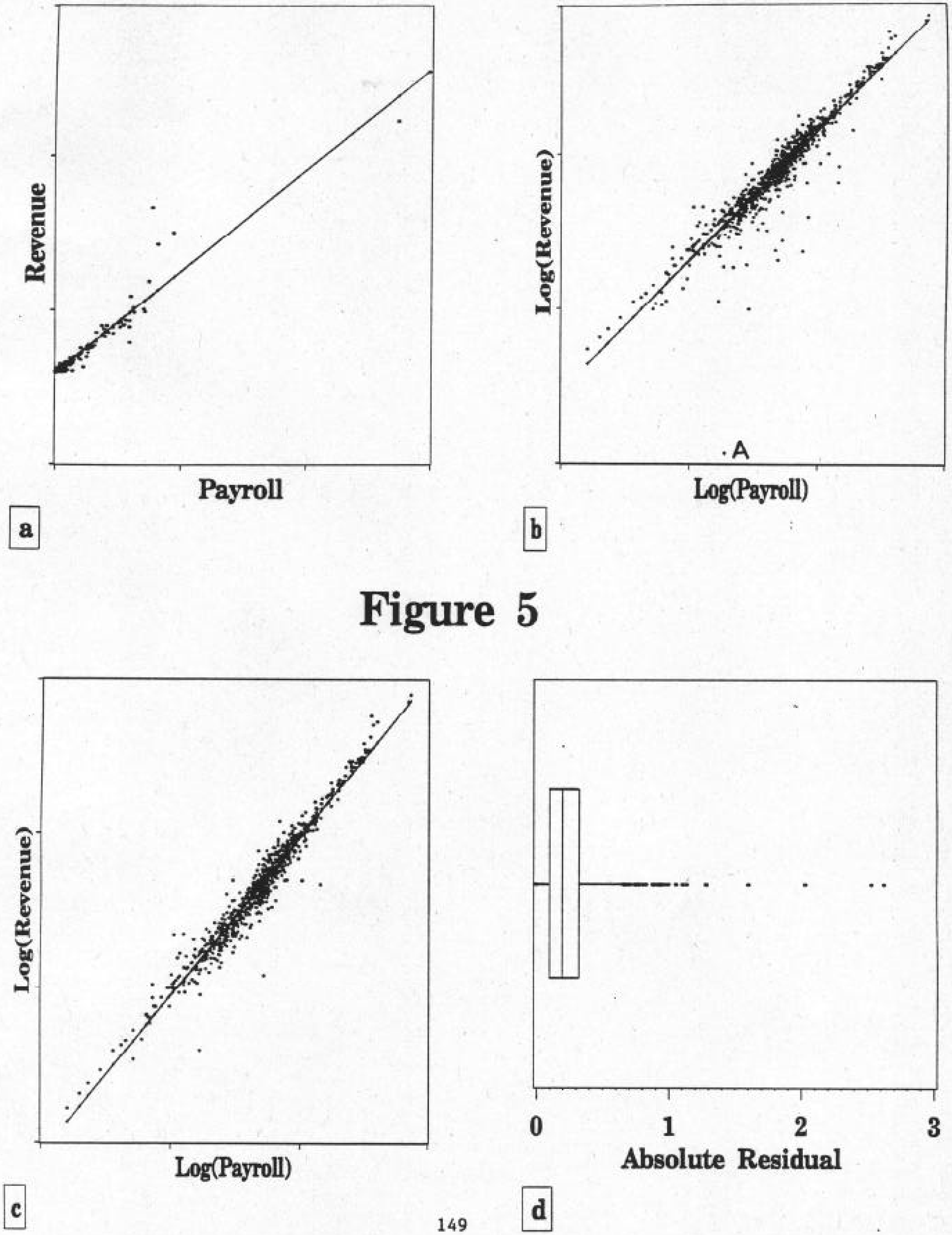

Revenue
Payroll
a
Log(Revenue)
A
Log(Payroll)
b
Log(Revenue)
Log(Payroll)
c
0
1
2
3
Absolute Residual
d

Figure 5

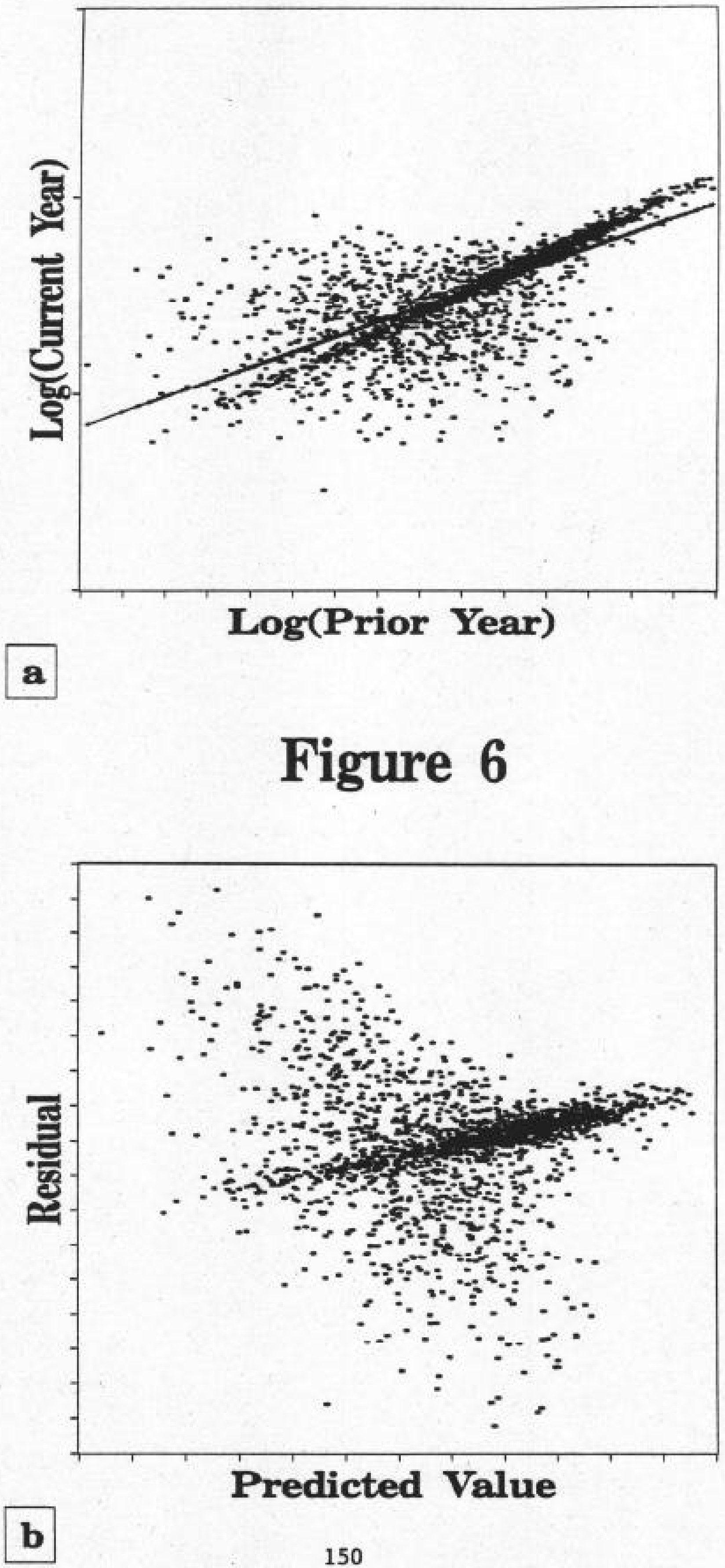
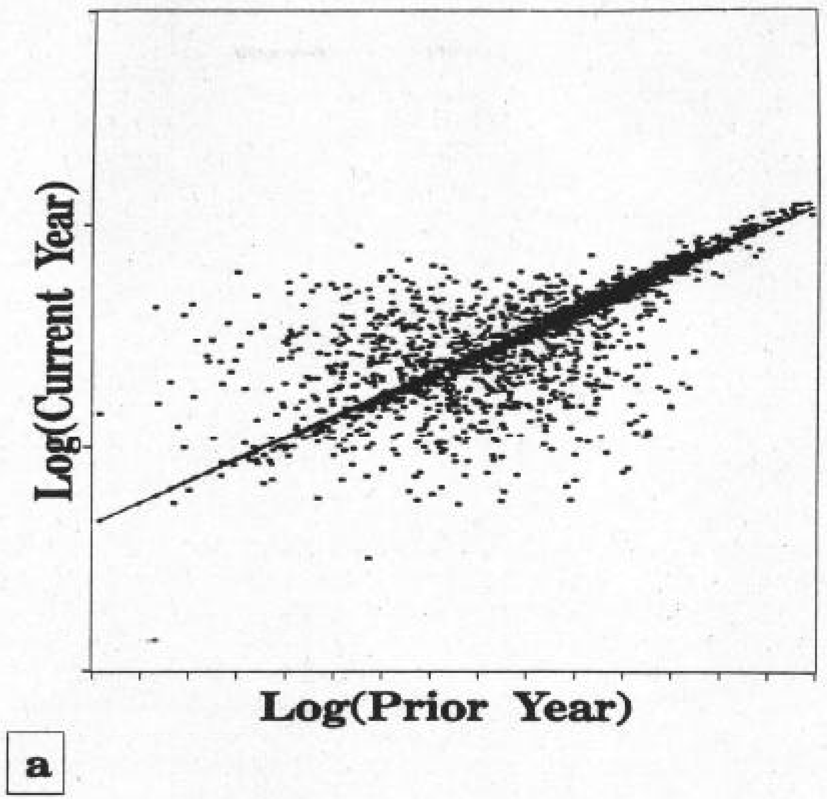
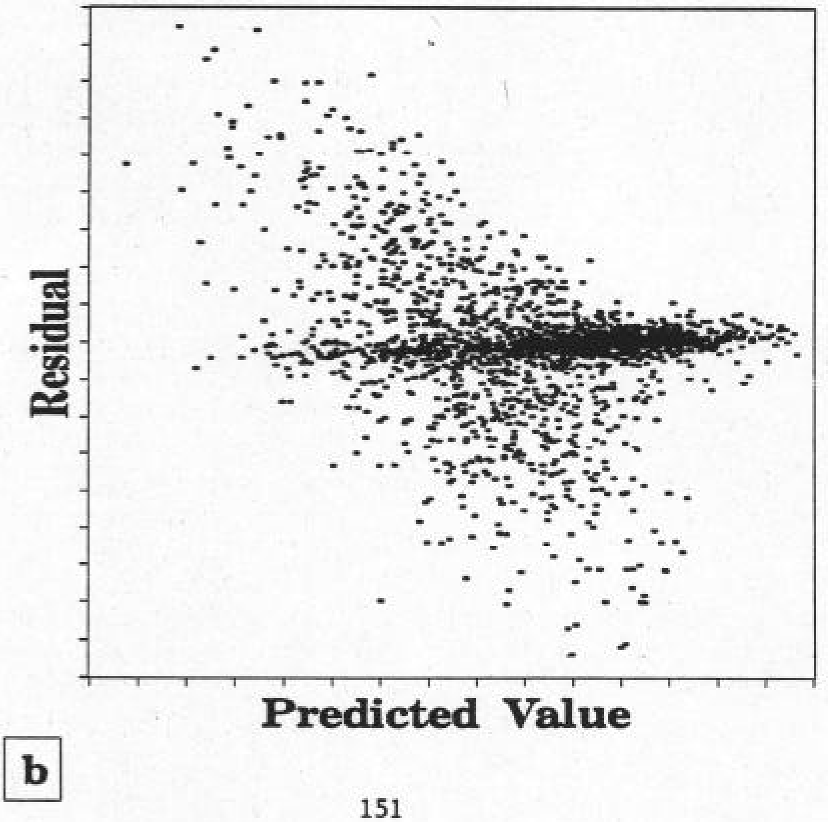

Log(Current Year)
Log(Prior Year)
a
Residual
Predicted Value
b

Figure 6



Log(Current Year)
Log(Prior Year)
a

Figure 7



Residual
Predicted Value
b

TIME SERIES AND CROSS SECTION EDITS WITH APPLICATIONS TO FEDERAL RESERVE DEPOSIT REPORTS

David A. Pierce and Laura Bauer Gillis[1]
Federal Reserve Board

ABSTRACT

Currently data from the major deposit reports submitted by commercial banks to the Federal Reserve System are edited by comparing the incoming value for a variable to that variable's value for the previous week, using a set of published *tolerances*. The previous value represents an estimate or forecast of what the current value would be in the absence of error or unusual circumstance. This paper investigates two generalizations of this editing method, which both involve incorporating information beyond that contained in the previous week's value. One of these is to base this estimate on the item values from a *cross section* of similar institutions in the current time period which have already reported, and the other is to calculate a forecast based on the *time series* of past values of the item. A composite estimate combining these two methods is also presented. Edit simulations are performed to measure the improvement from this approach (in terms of fewer edit exceptions which are correct and/or increased detection of errors), which is found to be substantial for some items and size groups. Efforts thus far to implement these enhancements are described, and possible further generalizations are mentioned.

1. INTRODUCTION AND SUMMARY

Data for the U.S. Money Supply are regularly transmitted to the Federal Reserve System by commercial banks and other financial institutions at weekly and other intervals. A major vehicle for this transmission is the "Report of Transactions Accounts, Other Deposits and Vault Cash", or simply the "Report of Deposits", on which banks and other financial institutions report weekly data for 25 deposit categories and related items. Based on these data and on similar information contained in other reports, the money supply measures are constructed and reserve requirements are maintained.

The money and reserves figures are important both as barometers of economic activity and in enabling the Federal Reserve to perform its economic stabilization and bank regulatory functions, and it is essential that the data submitted on the Report of Deposits and other reports be reliable and of high quality. To ensure their accuracy, all such data are subjected to numerical edits to detect unusual or deviant values. These edits are to two general types, *validity* edits to ensure that adding-up and other logical constraints are satisfied, and *quality* edits based on statistical or distributional aspects of the data.

[1] The authors are respectively Senior Statistician and Statistician, Division of Research & Statistics, Federal Reserve Board, Washington, DC 20551. The valuable assistance of Mia Johnson is gratefully acknowledged. Any views expressed do not necessarily reflect those of the Federal Reserve System.

The most commonly used quality edit involves the comparison of an incoming weekly figure to the previous value of that variable (in both dollar and percentage terms), using a tolerance band constructed about that value. The *tolerances*, or half-widths of the tolerance bands, are determined from previous estimates of the variable's distribution, in particular measures of spread, and are published in a Technical Memorandum or "Tech Memo"[2]. An edit "exception" occurs if the incoming value falls outside this tolerance band; when this happens, the reporting bank or other institution may be contacted for verification or correction. All tolerance-table comparisons are made (and edit exceptions generated) by machine, whereas the decision to contact the respondent is made by data analysts. The editing is done at both the Federal Reserve Board and the 12 Federal Reserve Banks.

Edits are in essence hypothesis tests, and errors of both kinds can occur. A major task in setting edit tolerances is to ensure adequate sensitivity without generating unnecessarily large quantities of "false positive" edit exceptions. It is because of the large number of exceptions currently generated that editing at both the Reserve Banks and the Board is currently quite labor intensive. All exceptions are reviewed by data analysts who must decide which are to be referred to the respondent institution for verification or revision. At the same time, a large majority of the data errors are not caught by these edits, based on the historical record of revisions submitted by respondents (they may be detected by other edits at a later date). There is consequently a need both to increase the sensitivity of the edits and to streamline the data editing process.

The previous value of the variable being edited, to which the tolerances are applied, in effect represents an estimate or forecast of the current figure in the absence of error or unusual circumstance. By basing this forecast or estimate on information beyond that contained in the previous week's value for that variable or item, we obtain the generalizations of the current editing method that are investigated in this paper. One generalization is to base this estimate on the item values from a *cross section* of similar institutions in the current time period which have already reported, intending to capture economic, institutional or calendar movements which tend to affect similar respondents in a similar manner. The other is to calculate a forecast based on the *time series* of past values of the item for that respondent, including possibly last month's or last year's figures in addition to the one for last week as in the current procedure. A composite estimate combining these two methods is also investigated, the idea being that each method may incorporate information not captured by the other. (We also generated a composite of the cross section and current edits).

The paper's focus is on the data submitted on the Report of Deposits, also known as the Edited Data Deposits System (EDDS) Report. We investigated four of the more important items on this report, total transactions deposits, total savings deposits, and large and small time deposits. The study was motivated by the desire for greater automation in the Federal Reserve Board's Division of Information Resources Management, which carries out the edits. The improvements resulting from the study are being incorporated into a new software package called DEEP (Distributed EDDS Editing Project), for interactive editing on the PC.

Our results vary greatly according to item, entity type (e.g. commercial bank, credit union, etc.), and the amount of data in an institution group -- the latter being important for reliable cross-

[2] "Processing Procedures for the Report of Transaction Accounts, Other Deposits and Vault Cash (FR2900), Technical Memorandum No. 16, Publications Section, Federal Reserve Board (December 1993).

section estimates. In most cases we find that, with sufficient data, the cross section approach is as reliable as the current editing procedure. For total transactions deposits almost uniformly, and for total savings deposits for most commercial bank categories, time series modelling plays a significant role in the edits.

The following section of the paper discusses in greater detail the methodology underlying the different data editing approaches investigated. Section 3 then describes a set of edit simulations we performed with each of the five types of edits studied, and presents the results of these. Based on the simulation results, we provided a set of recommendations for experimental edits for DEEP, for each entity type and item, which have recently become operational.

2. METHODOLOGY

Given a variable or item of interest, many data editing procedures can be characterized as first generating a forecast (a point estimate) of the incoming value for that item, next applying a tolerance to the forecast to form a tolerance interval (an interval estimate) for the incoming value, and then flagging that value if it is outside the tolerance interval. In the current editing framework, that forecast is taken to be the previous week's item value, and the tolerance is as given by the Tech Memo (footnote 2). In this section the two generalizations to the forecast noted in Section 1 are presented, along with composite procedures, after first describing the data and framework used.

2.1 Choice of Items and Statistical Form

The current approach to editing data from financial institutions is to subdivide them into homogeneous "cells", which are combinations of an institution's size group, entity type, geographic location. There are six size groups for commercial banks and a smaller number of size groups for each of the other entity types, which are credit unions, S&Ls, savings banks, agencies and branches of foreign banks, and Edge and Agreement Corporations. The geographic locations are defined in terms of 12 Federal Reserve districts.

There are thus a great many edit cells, and to make our task manageable, and to achieve comparability with the current edits, we have simplified this study in the following ways:

1. Staying with the *same cells* of the current EDDS edits. This will facilitate assessing the effects of the cross section estimates, model forecasts, and composite procedures. We recognize that more sophisticated groupings into cells may enhance the performance of the edits and plan to work with these in the future. Also we have eliminated all acquisitions and mergers from the institutions studied and have placed "credit-card banks" in a separate group.

2. Maintaining the *same tolerance widths* as currently (applied, however, to the time series / cross section estimates that we generate, as well as to the most recent value as currently done). This may at first seem unnecessary, since standard deviations, percentiles, and other aspects of the distribution can be determined from either the cross section data or the historical model. However, such calculations can sometimes be unreliable, especially with cross sections without at least several hundred institutions

in a group, as we are working with the extremes of distributions. And as with the cells themselves, keeping the current cell tolerance-interval widths facilitates comparisons among procedures.

We have also confined our attention in this study to the smaller institutions ("Priority-3" or P-3 institutions), where there may be the greatest potential for human resource savings from this approach. (Essentially this excludes the largest three size groups for commercial banks and a portion of the largest size group for other entity types). For these institutions, we have examined the following items:

Total transactions deposits	Large time deposits
Savings deposits	Small time deposits.

Current EDDS editing is performed with both dollar and percentage changes of the item being edited, with both required to exceed tolerances ("and" condition) for an exception to occur. The modifications outlined in this report are only for percentage changes; the Tech Memo tolerances continue to be applied to the dollar changes. There are several reasons for choosing percentage changes as the focus. Since they are used in current edits, the present edit cells and tolerances can be employed, and comparisons with current procedures can be made. They (or their annualized versions, growth rates) are also used in other analyses, such as with the Small Bank Sample of early reporting institutions. They are more homogeneous than dollar changes among different sized institutions, so that fewer edit groupings should eventually be needed. Percentage changes were found to be more sensitive to reporting and other errors than ratios to other items such as total deposits, which change with the denominator as well as the numerator and moreover present difficulty when the denominator was zero.

2.2 Cross Section Edits

Period-to-period edits compare an institution's current value for an item to the previous period's value. However, useful additional information may be contained in the current values of that item for other institutions that are similar to the one being edited. For example, if most of the institutions in a group experience a surge in large time deposits in a given week, then it would probably be inaccurate to list them as exceptions simply because they were outside the EDDS tolerances. Conversely, a very small change that week in large time deposits for a particular institution in that group may be suspicious even though current period-to-period tolerances would not be exceeded.

Cross section edits are carried out by examining the distribution of values (here, of percentage changes) for institutions within a homogeneous group, and listing as exceptions any values that were unusual compared to that distribution. Ordinarily one would calculate the mean and standard deviation of the percentage changes and flag those that were farther away from the mean than (say) two or three standard deviations; but in the present study we modified this set-up in two ways. First, because extreme values (the ones we hope to detect) would themselves influence the mean to which they would be compared, we "trimmed" the mean by eliminating the largest and smallest 5% of the values before calculating the estimated mean. Second, more observations are required to form a reliable estimate of the standard deviation than of the mean, and since most of the cells or groupings of institutions were too small for this, we chose to use multiples of the current EDDS tolerances as proxies for the standard deviations. As noted earlier, an additional advantage of this practice is to facilitate comparisons with the current edits.

One difficulty in using a cross section edit is that the data for an editing group need to be available in order to calculate such quantities as the average percentage change for that group. But the data for Priority-3 institutions are not due at the Board until nine days after the as-of date; and since timely estimates of the monetary aggregates and required reserves are needed, the editing process cannot be postponed this long. Our solution to this is to wait until a large enough fraction of the institutions have reported, and to form the distributional estimates (the trimmed means in this case) from the data available at that time.

For the EDDS data, more than half of the P-3 institutions' records are received by the Federal Reserve Board on the Thursday night following the as-of date (the previous Monday, on which the statement week ends), with the majority of those outstanding arriving by Friday night and the few remaining ones by the following Wednesday. For this study it was therefore decided to start the cross section editing on Friday morning, although work in progress is comparing this with the alternative of beginning on Monday morning. In either case, the trimmed mean estimates initially formed are not modified when more institutions have reported, in order not to confuse the editing process.

Some of the editing cells contain only a small number of respondents (and an even smaller number reporting by Friday), so that the estimated mean for those cells may not be very reliable. We required a minimum of 50 available observations in order to use the cross section estimate by itself. If the number of available observations is less than 50 but at least 20, a composite (see Sec. 2.4) of that estimate and the previous week's value for the institution is employed, and with less than 20 the previous week's value alone is used.

The cross section edit is performed by comparing the deviation between the observed and the estimated percentage changes to the current EDDS edit tolerance for the item. As noted earlier, if the percentage-change condition is violated, then a second comparison of the magnitude of the dollar change versus its tolerance is performed, and the item is flagged only if both sets of tolerances are exceeded. An exception to this is that, as is done with the current edits, when the item changes from zero to a nonzero value or vice versa, the current dollar-change edit tolerances are applied without any adjustment.

2.3 Time Series Edits

These edits are based on time series models, which predict or explain an item's present value in terms of its past history. This usually involves the immediately previous value, on which the current edits are based, and often additional values as well, such as last year's. To the extent that these more distant values are important in predicting the incoming value, more sensitive edits should result from taking them into account.

Editing using a time series model for generating forecasts of percentage changes implies that a historical relationship exists between the item and its previous values. The "random walk" model is a time series model in which the best forecast of the current value is simply last week's value. Thus, the random walk model is implied by the current period-to-period change edits, which take last week's value as the current-period forecast around which the tolerances are applied. More complicated time series models yield forecasts which are weighted averages of several past values of the percentage change.

We first investigated the fitting of time series models for each institution separately. Some institutions' data fit the models quite well, with reductions in the standard deviation of the forecast errors (a key to the effectiveness of tolerances of a given width) of 50% or more, while other institutions exhibited only weak fits, or only the random walk behavior that the current editing framework already captures. Although fitting individual models is the preferable method for forecasting, it was not feasible to maintain over 8000 models for each item edited within the DEEP framework - at least not at this time. Thus, at this stage and for the P-3 institutions, a single time series model was fit to each editing cell's aggregate, and the coefficients from that estimated model were used to obtain an individual bank's forecast using its own previous values. While the benefits of time series modelling are reduced by doing this, the method can be easily implemented, and updated when necessary. Another constraint at present is that, because of data storage limitations, we only utilized terms in the model at lags of 1, 2, 3, 52 and 53 weeks, thus capturing nearby effects and annual seasonal influences but not, say, monthly or quarterly effects.

As an example of the model-fitting results, Table 1 provides information on time series models fit to cell aggregates of Total Transactions Deposits for three of the editing cells. Notice the highly statistically significant seasonal effect (lag 52, and in some cases lag 53). The strength of the fit declines going down the page, with the third one (Edges & Agreements, a root MSE reduction of only 9.2%) being not much different from the random walk model underlying current edits. On the other hand the results suggest that model-based editing may be valuable for certain commercial bank cells, for total transactions.

As with cross section edits, the deviation between the actual percentage change and the forecasted change from the time series model is compared to the edit tolerances. A tolerance exceedance both here and on the dollar change (also using current EDDS tolerances) triggers an edit exception for the record.

2.4 *Composite Edits*

The cross-section and time series edits are based on different sets of information, past values of the institution being edited and present values of similar institutions. Thus a forecast which combined these two estimates, thereby utilizing both sources of information, may be more accurate than either one separately, and edits derived from such forecasts correspondingly more sensitive.

For a given institution (e.g. bank) and a given item, if T denotes a time-series estimate (forecast) for a given week, C represents a cross-section estimate, and A the actual value that is reported, then the <u>composite estimate</u> is a weighted average of T and C which is of the form

$$\omega T + (1-\omega)C.$$

The weights ω and $1-\omega$ depend on the relative sizes and the correlation between the estimation / forecast errors of T and C. If these errors are given by

$$ET = A - T \text{ and } EC = A - C\,,$$

then

$$\omega = [\mathrm{Var}(EC) - \mathrm{Cov}(ET.EC)] / \mathrm{Var}(ET\text{-}EC)\,.$$

A composite forecast is thus a weighted average of individual component forecasts where the relative weights are chosen to minimize the sum of the squared forecast or estimation errors, and where the sum of the weights is one.

Using past data, we investigated a composite estimate of the cross section and the time series forecasts, denoted "CSTS", for each editing cell and each item. The composite forecast defaults to the time series forecast with fewer than 20 available observations in the cell average. (With exactly 20 and using the 90% trim, 18 observed changes would be used in the cell estimate).

The other type of composite edit we considered combines the cross section and the random walk forecasts (CSRW). We employed this edit when a CS edit was indicated but the sample size - - the number of observations available on Friday morning when the cell means are formed -- was insufficient (less than 50) to obtain an adequately reliable cross section estimate. For very small sample sizes (less than 20), our procedure is to revert to the use of only the RW edit.

3. MODELLING AND SIMULATIONS

To examine the relative performance of different types of edits, we conducted simulations of these edits over the 1991-92 time period. For each cell (choice of item, entity type, size group and geographic region), we performed five sets of simulations, corresponding to the different types of edits under consideration: current (random walk), cross section, time series, cross section/time series composite, and cross section/random walk composite.

3.1 Simulation Procedure

Data preparation was a time consuming task. First, all Priority-3 reporters' weekly average data were compiled for the period from January 1986 through December 1992. While the edits were simulated only for the most recent two years, the additional data were used for fitting time series models with potential annual patterns. To avoid distortions, we eliminated all banks involved in mergers during this period. We next partitioned the data set into the editing groups or cells. We found that not all cells had a sufficient number of reporters to fit a model or to obtain reliable cross section estimates, and so some of them were combined. For commercial banks of size group 3 (total deposits between $1B+ and $3B), there were too few P-3 reporters to employ any of the new approaches. In addition, we added an editing category for known credit card banks. In total there were 40 edit cells, 37 of which were involved in the simulations.

Once the data were prepared, time series models were fit to the percentage changes in each cell's aggregate, as described in Section 2.3. Using the fitted model for a cell, predicted values for the last two years were generated for each institution in the cell. (Although forecasted values of the percentage change were generated for all periods, those in which a change of zero to a value or a value to zero were edited using the current special tolerances). Both the model-based and the zero-valued random walk forecasts were assigned to each observation in the cell. The 10% trimmed mean of the percentage changes was also calculated for each cell and each week of the two year simulation period, for use in the cross section edits. (Since the cross section simulation employed all the data within a cell to calculate the current-period forecast, rather than the available data as of Friday morning when editing begins, the simulated results will differ from those in practice). In order to generate the two composite forecasts, the prediction errors from the original three forecasts were computed and the formulas in Section 2.4 applied by institution. A cell root mean square prediction error (RMSE) was also computed.

Since the composite forecast combines the component forecasts in such a way as to minimize the sum of the squared prediction errors, we chose to estimate the appropriate weights for each bank in a cell and then to average those weights over the cell in order to obtain the composite for editing. Since the composite is a weighted average of the individual forecasts, the sum of the weights must equal one. For some institutions, where the prediction errors were very highly correlated between methods, we obtained pairs of weights with one value less than zero and the other greater than one. Evidently it only requires a small number of observations away from that correlation structure to cause such disproportionate weights. In calculating the average pair of composite weights for each cell, therefore, we first screened out those sets of weights not within the (0,1) range. After the two composite weighting schemes were determined for each cell, the mean square prediction errors were computed for these two forecasts as well.

For each of the five edit methods, Table 2 presents the root mean square prediction errors and composite weights for the commercial bank cells for total transactions and total savings, and Table 3 presents the same information for the other entity cells, for total transactions. We anticipate the method with the smallest forecasting error to have the best potential as an edit, but until our tolerances are better tailored to the actual editing method, this potential may not be realized.

To apply the edits, we first looked for percentage changes that differed from the forecasted percentage changes by more than the appropriate tolerance (whether taken from the Tech Memo or generated as described in this paragraph), and for those ascertaining whether the dollar change tolerance was also exceeded. Since total savings and large time deposits are currently edited items, their current tolerances can be used. However, for total transactions and small time deposits, current tolerances do not exist. We therefore generated tolerances in a manner similar to that used for the creation of the current ones. This involved iterative steps with the intent of flagging approximately 0.3% of the observations per cell on average (the maximum percentage of observations flagged using current editing methods for other items, for the year 1991). Using the components of total transactions and items that were related to small time, such as total and large time, we first compiled a range of feasible values for the tolerances. We then examined where these values occurred on the distribution of percentage changes over each cell for the two-year period. Given a reasonable proportion of the changes exceeding the initial values, we then examined the dollar change distribution for the subset of percentage change exceptions. Appropriate percentiles of this distribution were then determined to obtain the expected 0.3% edit failures under the current random walk model. These percentiles became the dollar change tolerances.

Once all the forecasts and tolerances were in place, the editing experience for the 1991-92 period was simulated for each of the five forecast methods. For each method we observed which observations were flagged as edit exceptions. Then based on a history of weekly revisions to the EDDS file maintained by the Federal Reserve's Statistical Services branch, we were able to determine the rate of type I and type II errors for each method. [A type I error (a "false positive") refers to an item that was flagged but not in error, or at least not revised. A type II error occurs when an item is not flagged but is erroneous (as evidenced by a later revision)].

3.2 Simulation Results

For reference in this section, Table 4 shows our recommended edits based on these simulations. As mentioned in Section 1, these are currently being implemented as part of the Federal Reserve Board's DEEP editing software. In Table 4, the left column lists the entities (with the included size groups in parentheses), followed by the chosen edit for each item.

Turning to the results on which this table is based, Table 5 summarizes the editing simulations for commercial banks; those for other entity types were similar and are given in an earlier report[3]. To assess the magnitudes and the implications of errors caught and errors missed by the editing schemes, the tables break down these errors in termr of their size (i.e. the size of the revision--we assume, however accurately, that revised data are correct and the revision is the error in the unrevised data). Each section of these tables compares the current (random walk) method with an alternative editing strategy. It is clear from these simulations that there is room for improvement, especially regarding the type II error probabilities, which range from 98% to 99%. And although the type I error probabilities appear small, the number of flagged items that are not in error is quite large (between 87% and 94%).

Wherever the fitted time series model indicated a potentially substantial payoff relative to the random walk model (as in the first model in Table 1), the time series edit tended to be the most accurate, yielding the smallest number of edit exceptions and with fewer errors missed that were captured by other methods than vice versa. The reduction in the number of edit exceptions was not as great for the CS and CSTS composite methods, but often the composite method caused less of an increase in the type II error probability. The CS and the CSRW composite often mimicked the current RW results. Where there was doubt regarding the preferable edit method, we tended to favor the CS or CSRW -- even when the reduction in RMSE and the number of edit exceptions was small relative to the current (RW) method -- since cross section edits would allow possibly large shifts in behavior for a given week to be incorporated into the editing norm, and the DEEP software is well-suited to this type of edit. Also, we gave some preference to a uniformity of editing method across related cells (e.g. adjacent size groups within an FR region, or like size groups between regions).

For commercial banks, the alternative edits on the whole did quite well. The time series edits for total transactions and total savings were effective in reducing the total number of exceptions while missing only 3 small revisions and actually finding an additional error of over $25M.[4] For the other entity types, total transactions was the only item that allowed for an alternative other than the CSRW method (CSRW was selected for these entity types in place of CS in order to accommodate smaller sample sizes in the preliminary data). Those credit unions and savings institutions which would have more activity in transactions accounts than the other entity types, do exhibit cyclical patterns which the time series model was able to capture (See Table 3.A). Agencies and branches also exhibited

[3] "Editing in DEEP: Utilizing Time Series and Cross Section Information", Laura Bauer Gillis and David A. Pierce, Federal Reserve Board, 1993 (preliminary report). Available from the authors.

[4] This revision was generated either by an outside source or by an edit of another report that is not being considered here. This occurrence brings to light that some errors are detected by other sources - not the Reserve Banks or the Board. What we gain from this additional edit exception an earlier detection of the error; it would not necessarily go undetected permanently.

improved editing results with the CSTS method. As mentioned, this combination of alternative strategies yielded an 11% reduction in both the type I error probability and the number of edit exceptions, with only a very slight increase in the type II error likelihood (about 0.1%).

All of these results are based on simulations using 1991 and 1992 EDDS data. Any errors caught before the data arrived at the Board are not reflected in these data, nor are errors undetected by Banks or Board that do not show up in the revision files. And as previously mentioned, the other factor to be monitored is the use of preliminary data in cross section estimates of the mean percentage change. Depending on how and where the preliminary data fall in the distribution of all percentage changes for an item, the operational results based on the CS, CSTS, or CSRW methods may differ significantly from what is expected based on the simulation results. The data availability and timing issue for cross section estimates is currently being studied.

This investigation is still in progress, and further generalizations of the work are underway or planned. Among these are examining time series models with regression components to account for such phenomena as tax dates, calendar effects or related variables, alternative groupings of the data according to size or geographic region, modelling larger banks individually, and examining additional items or variables.

Table 1. Percentage Change Models for Total Transactions Aggregates, Selected Editing Cells

------------------Cell = CB, Size Group 4, Region I------------------

Root MSE(orig.) = 0.0383 Root MSE(model) = 0.0211
Reduction in Root MSE = 44.9%

Variable	Parameter Estimate	Standard Error	T-stat	p-value
TRN$_{t-1}$	-0.4349	0.0483	-9.005	0.0001
TRN$_{t-2}$	-0.0341	0.0329	-1.039	0.2996
TRN$_{t-3}$	-0.1510	0.0338	-4.467	0.0001
TRN$_{t-52}$	0.6494	0.0318	20.391	0.0001
TRN$_{t-53}$	0.4668	0.0440	10.606	0.0001

---------------Cell = CU, Size Group 2, Regions II&III---------------

Root MSE(orig.) = 0.1067 Root MSE(model)=0.0809
Reduction in Root MSE = 24.2%

Variable	Parameter Estimate	Standard Error	T-stat	p-value
TRN$_{t-1}$	-0.2450	0.0546	-4.486	0.0001
TRN$_{t-2}$	-0.1160	0.0474	-2.444	0.0151
TRN$_{t-3}$	-0.2200	0.0486	-4.525	0.0001
TRN$_{t-52}$	0.4922	0.0477	10.312	0.0001
TRN$_{t-53}$	0.1866	0.0533	3.498	0.0005

----------------------------Cell = EA, All---------------------------

Root MSE(orig.) = 0.0564 Root MSE(model)=0.0512
Reduction in Root MSE = 9.2%

Variable	Parameter Estimate	Standard Error	T-stat	p-value
TRN$_{t-1}$	-0.3776	0.0569	-6.632	0.0001
TRN$_{t-2}$	-0.1547	0.0586	-2.642	0.0087
TRN$_{t-3}$	-0.0449	0.0553	-0.815	0.4181
TRN$_{t-52}$	0.2432	0.0524	4.638	0.0001
TRN$_{t-53}$	0.1057	0.0540	1.955	0.0514

Table 2. Root Mean Square Errors for Forecasts: Commercial Bank Cells

A. Total Transactions

	Root Mean Square Error					Weight of CS in Composite	
Cell	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
Region 1							
-Size 4	0.077	0.073	0.077	0.074	0.071	0.72	0.51
-Size 5	0.096	0.094	0.097	0.094	0.089	0.73	0.55
-Size 6	1.190	1.190	1.276	1.190	1.204	0.70	0.58
Region 2							
-Size 4	0.064	0.059	0.236	0.060	0.121	0.77	0.58
-Size 5	0.210	0.209	0.223	0.209	0.212	0.62	0.55
-Size 6	0.331	0.330	0.344	0.330	0.333	0.68	0.57
Region 3							
-Size 4	0.102	0.099	0.108	0.100	0.100	0.75	0.51
-Size 5	0.054	0.048	0.051	0.050	0.046	0.74	0.58
-Size 6	0.067	0.063	0.071	0.064	0.062	0.70	0.60

B. Total Savings

	Root Mean Square Error					Weight of CS in Composite	
Cell	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
Region 1							
-Size 4	0.042	0.042	0.045	0.042	0.042	0.64	0.73
-Size 5	0.054	0.054	0.056	0.054	0.054	0.64	0.67
-Size 6	0.048	0.048	0.055	0.048	0.048	0.60	0.72
Region 2							
-Size 4	0.038	0.038	0.099	0.038	0.043	0.65	0.76
-Size 5	0.235	0.234	0.244	0.234	0.236	0.64	0.64
-Size 6	0.055	0.055	0.067	0.055	0.055	0.64	0.66
Region 3							
-Size 4	0.051	0.051	0.998	0.051	0.274	0.68	0.74
-Size 5	0.041	0.040	0.041	0.040	0.040	0.63	0.66
-Size 6	0.055	0.055	0.065	0.055	0.055	0.61	0.75

Table 2. **Root Mean Square Errors for Forecasts: Commercial Bank Cells (Continued)**

C. Large Time	Root Mean Square Error					Weight of CS in Composite	
Cell	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
Region 1							
-Size 4	0.067	0.067	0.069	0.067	0.067	0.53	0.61
-Size 5	0.110	0.110	0.117	0.110	0.110	0.52	0.75
-Size 6	0.160	0.160	0.184	0.160	0.161	0.49	0.81
Region 2							
-Size 4	0.089	0.088	0.093	0.088	0.089	0.54	0.68
-Size 5	0.063	0.063	0.065	0.063	0.063	0.48	0.62
-Size 6	0.099	0.099	0.109	0.099	0.100	0.46	0.76
Region 3							
-Size 4	0.047	0.047	0.051	0.047	0.047	0.55	0.70
-Size 5	0.075	0.075	0.076	0.075	0.076	0.54	0.59
-Size 6	0.120	0.120	0.141	0.120	0.120	0.51	0.80

D. Small Time	Root Mean Square Error					Weight of CS in Composite	
Cell	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
Region 1							
-Size 4	0.064	0.064	0.098	0.064	0.068	0.58	0.70
-Size 5	0.143	0.143	0.156	0.143	0.144	0.52	0.70
-Size 6	0.110	0.110	2.274	0.110	0.409	0.55	0.83
Region 2							
-Size 4	0.468	0.468	0.516	0.468	0.470	0.59	0.79
-Size 5	1.363	1.363	1.420	1.363	1.369	0.61	0.68
-Size 6	0.034	0.034	0.036	0.034	0.034	0.58	0.77
Region 3							
-Size 4	0.062	0.062	0.068	0.062	0.063	0.61	0.67
-Size 5	0.017	0.017	0.018	0.017	0.017	0.58	0.65
-Size 6	0.063	0.063	0.072	0.063	0.063	0.57	0.80

Table 3. Root Mean Square Errors for Forecasts: Other Entity Types, Total Transactions

A. Agencies and Branches

	Root Mean Square Error					Weight of CS in Composite	
Cell	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
-All Regions							
Size 1	1.366	1.363	1.378	1.364	1.364	0.48	0.54
-Region 1							
Size 2	2.700	2.696	2.794	2.698	2.715	0.45	0.53
Size 3	5.061	5.061	5.974	5.061	5.198	0.38	0.62
-Region 2							
Size 2	2.248	2.240	2.406	2.245	2.317	0.38	0.35
Size 3	4.158	4.154	4.965	4.156	4.248	0.43	0.67
-Region 3							
Size 2	0.250	0.247	0.250	0.248	0.243	0.34	0.44
Size 3	4.300	4.289	5.416	4.295	4.544	0.42	0.58

B. Credit Unions

	Root Mean Square Error					Weight of CS in Composite	
Cell	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
-All Regions							
Size 1	0.106	0.073	0.075	0.076	0.069	0.75	0.53
-Region 1							
Size 2	0.093	0.069	0.059	0.075	0.059	0.57	0.37
Size 3	0.122	0.112	0.120	0.114	0.110	0.59	0.43
Size 4	0.084	0.075	0.078	0.077	0.073	0.57	0.45
-Regions 2 & 3							
Size 2	0.084	0.062	0.054	0.065	0.052	0.64	0.57
Size 3	0.099	0.080	0.078	0.083	0.073	0.63	0.37
Size 4	0.082	0.069	0.060	0.072	0.059	0.56	0.40

Table 3. Root Mean Square Errors for Forecasts: Other Entity Types, Total Transactions (Continued)

C. Edges and Agreements

	Root Mean Square Error					Weight of CS in Composite	
Cell	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
-ALL	17.21	17.20	18.94	17.20	17.72	0.44	0.47

D. Savings Institutions

	Root Mean Square Error					Weight of CS in Composite	
Cell	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
-Region 1							
Size 1	0.057	0.048	0.049	0.049	0.047	0.79	0.67
Size 2	0.187	0.185	0.193	0.185	0.185	0.74	0.57
Size 3	0.744	0.743	0.965	0.743	0.779	0.71	0.61
Size 4	0.627	0.626	0.645	0.626	0.627	0.66	0.63
-Regions 2 & 3							
Size 1	0.073	0.065	0.068	0.065	0.064	0.73	0.68
-Region 2							
Size 2	0.132	0.129	0.153	0.129	0.131	0.73	0.62
Size 3	0.077	0.072	0.079	0.073	0.072	0.78	0.66
Size 4	0.066	0.062	0.069	0.062	0.061	0.75	0.66
-Region 3							
Size 2	0.077	0.069	0.077	0.070	0.068	0.78	0.58
Size 3	0.309	0.308	0.766	0.308	0.377	0.72	0.66
Size 4	0.370	0.369	0.568	0.369	0.408	0.63	0.59
-Region 4							
Size 1	10.73	10.73	11.44	10.73	10.77	0.67	0.75

Table 4. Experimental Edits for DEEP

	Total Transactions	Total Savings	Large Time	Small Time
Commercial Banks (3,Ccd)	RW	RW	RW	RW
Commercial Banks (4,5,6)	CSTS	TS	CS	CS
Credit Unions (1,2,3,4)	TS	CSRW	CSRW	CSRW
S&Ls, Coops, Sbs (1,2,3,4)	ℜI ℜII-IV TS CSTS	CSRW	CSRW	CSRW
Agencies & Brs.(1,2,3)	CSTS	CSRW	CSRW	CSRW
Edges & Agr. (1,2)	CSRW	CSRW	CSRW	CSRW

The numbers in parentheses are the size groups, with "Ccd" denoting credit card banks. CB size groups 1 and 2 are omitted, as they are priority 1 and 2 institutions. ℜ denotes the FR Region, as in TM#16. The other entries in this table have the following explanations:

TS: The time-series model-based forecast, utilizing the institution's past percentage changes (of 1,2,3,52 and 53 weeks ago).

CS: The cross-section forecast, or estimate of the average percentage change over all the institutions in the editing group or cell. Uses only the data received by the Friday after the as-of date and is calculated as the 90% trimmed mean of the individual percentage changes in the cell.

CSTS: A weighted average of the TS and CS percentage-change forecasts, with statistically determined weights. When the number (n) of institutions in the group available on Friday for calculating the mean is less than 20, the weights are 1 and 0 (only the TS forecast is used).

RW: The forecast based on the "random walk" model, or the time series model giving a zero period-to-period change as the best forecast -- and is thus the implicit model underlying the current edits. This translates into a percentage-change forecast of zero.

CSRW: The forecast based on a composite of the CS and RW estimates of the percentage change, again depending on the number n of available observations in the cell. Thus:

if $n \geq 50$, use CS only;
if $20 \leq n < 50$, use weighted average of the CS and RW estimates;
if $n < 20$, use the RW estimate (zero percentage change forecast).

Table 5. Editing Simulation Results: Commercial Banks

A. Total Transactions

1. Random Walk (Current Editing)

Frequency/ Percent	Not Revised	<$5M	$5M <$10M	$10M < $25M	> $25M	Total
Not Flagged	557,166 97.76	9,732 1.71	791 0.14	508 0.09	168 0.03	568,365 99.73
Flagged	1,444 0.25	75 0.01	17 0.00	12 0.00	6 0.00	1,554 0.27
Total	558,610 98.01	9,807 1.72	808 0.14	520 0.09	174 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.26%
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 99.0%
Pr(Item not in error | Item Flagged) = 92.9%

2. Cross Section - Time Series Composite

Frequency/ Percent	Not Revised	<$5M	$5M <$10M	$10M < $25M	> $25M	Total
Not Flagged	557,326 97.78	9,743 1.71	792 0.14	509 0.09	167 0.03	568,537 99.76
Flagged	1,284 0.23	61 0.01	16 0.00	11 0.00	7 0.00	1,382 0.24
Total	558,610 98.01	9,807 1.72	808 0.14	520 0.09	174 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.23%
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 99.1%
Pr(Item not in error | Item Flagged) = 92.9%

Reduction in edit exceptions = 11.1%
Reduction in type I error probability = 11.5%
Increase in type II error probability = 0.1%

Table 5. Editing Simulation Results: Commercial Banks (Continued)

B. Total Savings

1. Random Walk (Current Editing)

Frequency/ Percent	Not Revised	<$5M	$5M <$10M	$10M < $25M	> $25M	Total
Not Flagged	557,547 97.83	8,772 1.54	723 0.13	375 0.07	181 0.03	567 598 99.59
Flagged	2,176 0.38	91 0.02	22 0.00	18 0.00	14 0.00	2,321 0.41
Total	559,723 98.21	8,863 1.56	745 0.13	393 0.07	195 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.39%
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 98.6%
Pr(Item not in error | Item Flagged) = 93.8%

2. Time Series

Frequency Percent	Not Revised	<$5M	$5M <$10M	$10M < $25M	> $25M	Total
Not Flagged	557,743 97.86	8,775 1.54	723 0.13	376 0.07	181 0.03	567,798 99.63
Flagged	1,980 0.35	88 0.02	22 0.00	17 0.00	14 0.00	2,121 0.37
Total	559,723 98.21	8,863 1.56	745 0.13	393 0.07	195 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.35%
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 98.6%
Pr(Item not in error | Item Flagged) = 93.4%

Reduction in edit exceptions = 9.8%
Reduction in type I error probability = 10.2%
Increase in type II error probability =. 0.0%

Table 5. Editing Simulation Results: Commercial Banks (Continued)

C. Large Time

1. Random Walk (Current Editing)

Frequency/ Percent	Not Revised	<$5M	$5M <$10M	$10M < $25M	> $25M	Total
Not Flagged	558,956 98.08	8,494 1.49	601 0.10	345 0.06	179 0.03	568,575 99.76
Flagged	1,248 0.22	68 0.01	19 0.00	8 0.00	1 0.00	1,344 0.24
Total	560,204 98.30	8,562 1.50	620 0.10	353 0.06	180 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.22%
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 99.0%
Pr(Item not in error | Item Flagged) = 92.8%

2. Cross Section

Frequency Percent	Not Revised	<$5M	$5M <$10M	$10M < $25M	> $25M	Total
Not Flagged	558,967 98.08	8,494 1.49	601 0.10	345 0.06	179 0.03	568,586 99.76
Flagged	1,237 0.22	68 0.01	19 0.00	8 0.00	1 0.00	1,333 0.24
Total	560,204 98.30	8,562 1.50	620 0.10	353 0.06	180 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.22%
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 99.0%
Pr(Item not in error | Item Flagged) = 92.8%

Reduction in edit exceptions = 0.8%
Reduction in type I error probability =. 0.0%
Increase in type II error probability = 0.0%

Table 5. Editing Simulation Results: Commercial Banks (Continued)

D. Small Time

1. Random Walk (Current Editing)

Frequency/ Percent	Not Revised	<$5M	$5M <$10M	$10M < $25M	> $25M	Total
Not Flagged	556,637 97.67	9,869 1.73	1,007 0.18	479 0.08	215 0.04	568,210 99.70
Flagged	1,496 0.26	117 0.02	42 0.01	36 0.01	18 0.00	1,709 0.30
Total	558,138 97.93	9,986 1.75	1049 0.18	515 0.09	233 0.05	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.27%
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 98.2%
Pr(Item not in error | Item Flagged) = 87.5%

2. Cross Section

Frequency Percent	Not Revised	<$5M	$5M <$10M	$10M < $25M	> $25M	Total
Not Flagged	556,637 97.67	9,869 1.73	1,008 0.18	479 0.08	215 0.04	568,211 99.70
Flagged	1,496 0.26	117 0.02	41 0.01	36 0.01	18 0.00	1,708 0.30
Total	558,138 97.93	9,986 1.75	1049 0.18	515 0.09	233 0.05	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.27%
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 98.2%
Pr(Item not in error | Item Flagged) = 87.6%

Reduction in edit exceptions = 0.0%
Reduction in type I error probability = 0.0%
Increase in type II error probability = 0.0%

DISCUSSION

Sandra A. West
U.S. Bureau of Labor Statistics

Let me first commend both sets of authors on very interesting and informative papers. Let me start with the David Pierce and Laura Bauer Gillis paper, "Time Series and Cross Section Edits with Applications to Federal Reserve Deposit Reports."

I enjoyed this paper very much; it was nice to see editing formally enter the realm of statistical inference. One might think of Imputation as point estimation, and editing as interval estimation--or perhaps as multiple imputation. I'd like to focus on one of the editing techniques in terms of imputation, but first let me briefly summarize the study.

In the paper there are:

5 methods for editing percent changes

1. Assuming no change from last week--current method-random walk, **RW**,---would be called Carry Over in nonresponse literature.
2. Using a cross section, **CS**, of similar respondents in the current time period which have already reported. Underlying assumption here that the previous time period values are available. (For surveys that do have nonresponse, only those entities that have reported in both time periods would be used.)
3. Using a time series, **TS**, of the past values of the respondent.
4. Composite of 1 & 2, **CSRW**.
5. Composite of 2 & 3, **CSTS**.

Several entity types-Respondents

Commercial Banks (There were two categories of this type.)
Agencies and Branches
Credit Unions
Edges and Agreements
Savings Institutions

Although there are 25 variables collected, the following 4 were studied.

Variables Collected from each Respondent

Total Transactions
Total Saving
Large Time
Small Time

Edits are performed weekly for a span of two years

Edits are performed in homogeneous cells which are combinations respondents' size, type and geographic location.

I'd like to discuss the cross sectional estimator, CS, since I've had some experience with a similar one using BLS data in terms of imputation. First, I need some notation. In a given cell, let

$Y_{t,i}$ = level of item for entity i at time t.

$\hat{Y}_{t,i}$ = predicted level for entity i at time t.

Editing is performed with percentage changes of the item; that is,

$$D_i = \frac{Y_{t,i} - Y_{(t-1),i}}{Y_{(t-1),i}}$$

(and for the current method the actual changes are also required to be in the tolerance interval).

For the CS edits, the empirical distribution is formed for the percentage changes, and the trimmed mean is computed, where 5% of each tail is trimmed. (Later I will say something about the trimmed mean.) Multiples of the current tolerances are used for proxies for the standard deviation. We could write the trimmed mean as

$$\overline{D} = \frac{1}{m}\sum_{i\in M}\frac{Y_{t,i} - Y_{(t-1),i}}{Y_{(t-1),i}} = \frac{1}{m}\sum_{i\in M}\frac{Y_{t,i}}{Y_{(t-1),i}} - 1$$

where M denotes the set of entities for which the percentage changes are in the middle 90% of the distribution, and m is the number of elements in M.

Using this technique, I'd like to come up with an imputation method. If we let $\hat{Y}_{t,j}$ be the predicted value for the j th entity at time t, and we estimate the percentage change by the trimmed mean, we have the following formula:

$$\frac{\hat{Y}_{t,j} - Y_{(t-1),j}}{Y_{(t-1),j}} = \overline{D} = \frac{1}{m}\sum_{i\in M}\frac{Y_{t,i}}{Y_{(t-1),i}} - 1$$

which leads to $\hat{Y}_{t,j}$:

$$\hat{Y}_{t,j} = \overline{D}Y_{(t-1),j} + Y_{(t-1),j} = \frac{1}{m}\sum_{i\in M}\frac{Y_{t,i}}{Y_{(t-1),i}}\ Y_{(t-1),j}.$$

Thus, for the j th entity, we would use his previous time period value, adjust it by the mean ratio of those entities that have already responded, to obtain his current value.

Now looking at this from a regression point of view, consider:

$$Y_{t,j} = \beta Y_{(t-1),j} + \varepsilon_{t,j} \quad where\ \varepsilon_{t,j} \sim N(0, \sigma^2 Y^{\delta}_{(t-1),j})$$
$$for\ j \in M$$

Using a weighted least squares the predicted value is:

$$\hat{Y}_{t,j} = \hat{\beta} Y_{(t-1),j} \quad for\ j \in M$$

where

if $\delta = 1$
$$\hat{\beta} = \frac{\sum_{i \in M} Y_{t,i}}{\sum_{i \in M} Y_{(t-1),i}}$$

if $\delta = 2$
$$\hat{\beta} = \frac{1}{m} \sum_{i \in M} \frac{Y_{t,i}}{Y_{(t-1),i}} .$$

Both are unbiased estimators of β but the one that is more precise depends on the value of δ. $\delta = 2$ is what underlies the CS method.

Under many situations one can show that the sum of the ratios has better properties than the ratio of the sums. However, in a study we did at BLS considering alternative imputation methods, we found that the model with $\delta = 1$ did the best. This was a study involving employment and wage variables for establishments on the Universe Data Base. We investigated many different methods; among them was a generalized Bayesian model, which led to multiple imputation. We also considered a time series going back a year, but only the prior month in this simple model was needed.

$$Y_{t,j} = \beta Y_{(t-1),j} + \varepsilon_{t,j} \quad where\ \varepsilon_{t,j} \sim N(0, \sigma^2 Y_{(t-1),j})$$
$$for\ j \in M$$

Using a weighted least squares, the predicted value is

$$\hat{Y}_{t,j} = \hat{\beta} Y_{(t-1),j} \quad for\ j \in M$$

where

$$\hat{\beta} = \frac{\sum_{i \in M} Y_{t,i}}{\sum_{i \in M} Y_{(t-1),i}}.$$

The M in this case was a set of homogeneous establishments that had reported values in both time periods. For establishment j in time period t, $Y_{t,j}$ denoted, in various studies, the reported employment, the reported ln(wages), and the reported ln(wages/employment). I'm not sure which model would work best with Bank type data, but I think it's worth writing down the underlying models so they can be tested.

In imputation studies, we have a problem similar to one that exists with the CS method. In imputation, when modeling the respondents to predict for the nonrespondents, one hopes the nonrespondents are missing at random; that is, the nonresponse mechanism is ignorable. If this is not the case, it is a difficult problem to model the response mechanism. A similar situation arises with the CS method, in that the edit criteria are set by the early arrivals, and it is hoped that the respondents that are due late, behave in a similar fashion.

I have a couple of observations from the Tables. There were 24 edit groups, consisting of the 6 types of respondents and the 4 variables. Of these 24, for more than half (13), the recommended procedure is the composite of CS and RW. In most of the cases, the CS had the larger weight than RW. Clearly, some form of the CS technique is worth pursuing.

I note from Tables 5 and 6 that the probability of a type II error is very large, and in some situations the probability of a type I error is also large, but not as large--70's versus 90's. I would think that the type II error, not flagging a value when it's in error, is more important than the type I error, flagging a value when it's not in error. But from an analysts point of view, I can see that the type I error would be more important.

Now let me discuss the Julia Bienias, David Lassman, Scott Scheleur, and Howard Hogan Paper, "Improving Outlier Detection in Two Establishment Surveys." I also enjoyed this paper. First a brief summary of the paper.

This paper uses Exploratory Data Analysis (EDA) to improve the detection of outliers in the following two establishment surveys.

1. **The Annual Survey of Communication Services---**2000 firms
2. **The Monthly Wholesale Trade Survey---**7,000 firms, only 3,500 receive forms in a given month

Techniques discussed:

Box Plots
Scatter Plots
Transformations
Fitting: Ordinary Least Squares
Weighted Least Squares

As I mentioned earlier, in our imputation study for wages, we found that if we first transformed the data by the natural logarithm, and used a weighted least squares, the imputation improved.

In general, I believe EDA should be part of any outlier detection system. There is an extensive literature on testing for outliers. A number of popular procedures have difficulty when a sample may contain multiple outliers. Problems include **masking**, in which the presence of other outliers makes each outlier difficult to detect, and **swamping** in which the procedure tends to declare too many outliers. By using robust and resistant methods it is possible to minimize the effects of deviate observations. An example is given in the Hoaglin, Iglewicz, and Tukey, JASA, 1986 paper, "Performance of some Resistant Rules for Outlier Labeling". Here you have inner and outer fences with hinges formed by the lower and upper fourths. That is, using the lower and upper fourths, F_L and F_U, the inner rule labels as "outside" any observations below $F_L - 1.5(F_U - F_L)$ or above $F_U + 1.5(F_U - F_L)$. For the outer rule 1.5 is replaced by 3.

In comparing the two papers, I found that I would like to combine them. For example, in the cross sectional estimates of Pierce and Gillis, additional EDA techniques could be used. As an example, instead of trimmed means one might consider "adaptive trimmed means". Some "adaptive trimmed means" determine the amount of trimming according to a sample estimate of the tail heaviness of the underlying distribution. This is especially useful if the distributions are not symmetrical, which I assume is the case with bank data.

In closing, I 'd like to compliment the authors for very fine papers.

DISCUSSION

Brian V. Greenberg
U. S. Bureau of the Census

In this discussion we attempt to relate these two fine editing papers to the broader issues in data editing and highlight what one can learn from them.

1. Introduction--Role of Editing

Broadly speaking, there are two primary reasons for editing survey and census data. First, we would like to remove erroneous values from micro-data sets. A second, and related objective, is to ensure that we can generate meaningful estimates from reported data.

For some programs, there is an emphasis on actual micro-level data. For example, when one establishes a longitudinal data file or when a public-use micro-data file will be the primary survey data product. An example of a longitudinal micro-data file is described in the Pierce and Gillis paper and their edit activities focus on the underlying data set.

On the other hand, for some surveys there is a single estimate (or small number of estimates) produced from a survey, and the underlying data file is less important then the single estimate. The Census Bureau's monthly report of wholesale trade, as discussed in the Bienias, Lassman, Scheleur, and Hogan paper, is an example of such a survey.

In any event, data editing does not exist in a vacuum, and in designing and evaluating an edit system, one should be mindful of the survey's data collection and release objectives.

2. Editing Stages in Data Collection and Tabulation

There are typically three stages of data editing for a typical survey or census: (1) data entry edit, (2) automated batch edit of individual data records, and (3) review of summary tabulations.

In the data entry stage, editing often consist of rudimentary checks that only attempt to detect keying errors and major reporting problems. There are, however, data entry programs which have sophisticated and extensive data edit capabilities. For some surveys, editing in the data entry stage, including on-the-spot follow-up with respondents, serves as the primary edit activity.

Batch editing of individual data records, referred to as _micro-editing_, has been the mainstay of many large-scale survey and census programs. For some surveys, the automated program alters suspicious values, while for others the automated edit only flags suspicious values for analyst review and action. In addition, automated batch edit systems often impute for missing values.

After preliminary editing, data are tabulated and estimates are edited against prior time periods, against information from other sources, or against one another. The process of editing tabulation cells is often referred to as _macro-editing_. If a tabulation cell looks suspicious, it is reviewed and the individual micro-level records contributing to the cell are examined. Some programs have very sophisticated macro-edit systems while other macro-edit programs are essentially manual.

It is important to note that even though potential data errors are detected at the macro-level in a tabulation cell edit, problems are typically resolved at the micro-level.

After data are processed through automated edit programs, there is typically analyst review of large and/or important cases which often include direct follow-up with respondents. For large programs, there are not sufficient resources to review all records, therefore records are ranked by importance and those most important to a program are reviewed by analysts. The ranking process is often informal, however, research at Statistics Canada to formalize this process (referred to as selective editing) seems to have met with success for their Annual Survey of Manufactures.

All three edit stages come into play in virtually all survey programs. Emphasis on one stage or another is typically embedded in the edit strategy of each program, bearing in mind the proposed uses of survey products.

3. Edit Tolerances

At each stage of the editing process, edit tolerances are required to target individual records or tabulation cells for review and, if necessary, correction. In many respects, the two papers discussed here focus on deriving edit tolerances and we discuss them from this perspective.

There are several steps to deriving meaningful edit tolerances. First, one defines edit cells; that is, groups of respondents whose behavior is fairly similar with respect to the edit criteria. One typically wants cells to be small enough so that respondents can be relatively homogeneous yet large enough so that parameters are not unduly influenced by a few nontypical responses.

One can generate (explicitly or implicitly) an anticipated value for data fields or a relation between fields. The anticipated value may be based on data from the current or prior time periods and can be modeled based on all respondents in the edit cell. Tolerance limits are applied to target records which have unacceptable deviations from anticipated values. For example, anticipated values can be based on a regression line and the tolerance limits can reflect the allowable band of values about this line. Under alternative approaches to deriving tolerances one directly determines a range of acceptable data values and designates response combinations outside that range as edit failures.

Tolerance parameters are derived and applied at each of the three stages of the edit process: data entry edit, batch edit of individual records and tabulation cell edit.

User-friendly systems to support the review of data in the development of edit tolerances can be extremely valuable. It is in this light that the work of Bienias, Lassman, Scheleur, and Hogan can be viewed. The graphics techniques which they present are particularly important because they can help users organize and systemize information and share findings with others. Such systems provide analysts access to methodology not otherwise readily available to them.

4. Striking a Balance in Edit Tolerances and Review Criteria

If edit tolerances are too tight, excessive data may be altered or sent for analyst review. In the first case, edit programs can distort estimates and force data to conform to expectations. In the second, too many referrals place a major burden on analyst resources.

If edit bounds are too loose, erroneous data gets into the system. Such errors in data limit the usefulness of micro-data and may lead to unreliable estimates. Broadly speaking, parameters which are too loose deprive us of the chance to identify a source of nonsampling error.

After automated edit programs have applied tolerances and targeted records as suspicious, one would like to select the most significant problematic records for analyst review. For each survey, one needs a reliable criterion as to what is significant and what is not. The notion of significant depends a great deal on the proposed user of the survey data.

Recently, the phrase "over-editing" has come into vogue to refer to spending too much time and money on editing and/or changing too much data. I feel somewhat uncomfortable with this phrase because it is unfocused and gives a misleading impression. It seems to imply that if we edited less--perhaps had looser edit bounds or reviewed less micro-data--we would be editing better. In fact, we want to edit more cleverly, not necessarily more or less. That is, we would like to target fields for change and/or review where change is needed and not target fields for change and/or review when not needed.

A more useful formulation of the issue can be couched in terms of Type I or Type II error for the edit process, as was done in the Pierce-Gillis paper. Namely, for their purposes:

> A Type I error (a 'false positive') refers to an item that was flagged but was not in error, or at least not revised.
>
> A Type II error occurs when an item is not flagged but is erroneous (as evidenced by a later revision).

We can broaden their definition a little to say:

> Type I error refers to an item flagged for change or review but the time spent on it did not improve the data set or estimates for the survey.
>
> Type II error occurs when an erroneous value, which adversely affect the quality of the data set or survey estimates, is not flagged for change or review.

The last conference on Statistical Policy Working Papers of the Federal Committee on Statistical Methodology was held in March, 1991. At that conference, there was a session based on Working Paper #18, "Data Editing in Federal Agencies." In that report, the focus was on development of multipurpose systems, software design, and edit methodologies. There was little discussion of parameter development. Since then, it has been increasingly clear that good parameter development is crucial in all stages of editing. It is also clear that we need to give greater attention to the interplay between subject-matter staff and automated programs in the resolution of edit failures and in the design of edit tolerances.

Even the best edit methodology embedded in the finest system will perform poorly if there are bad parameters. In fact, the choice of edit tolerances has a major influence on the Type I and Type II error for editing. We certainly need more investigation to highlight what methods and tools work well for the design of edit tolerances and we need to examine and learn from clearly presented case studies.

The two papers under discussion do an excellent job in addressing these related issues.

5. Bienias, Lassman, Scheleur, and Hogan Paper

The authors illustrate graphic techniques used in the spirit of exploratory data analysis as tools for subject-matter specialists in deriving edit parameters. They also describe how the simultaneous review of survey data can introduce advantages over a case-by-case analysis of report forms.

Box plots were used to review and summarize information and directly contributed to parameter development for the Annual Survey of Communication Services (ASCS). In particular, the box plot for parameters based on the expense/revenue ratio illustrates this use.

Graphics were also used to help uncover similarities or differences between establishments. By examining residuals in the relation between revenue and payroll in ASCS, they decided to remove tax-exempt establishments from edit cells for revenue and payroll. That is, they were able to design a more effective edit cell for subsequent analysis, which they describe.

And finally, the graphics and exploratory data analysis led to a more suitable editing model for current to prior inventory on the Monthly Wholesale Trade Survey. In this usage, the techniques they employed allowed subject-matter analysts to experiment with different models and to select the model that they felt best represented the data.

An important theme of this paper was the interplay between subject-matter analyst expertise and the use of graphical methods. These tools can provide a guide for analysts and allow them to see the impact of proposed models. They can contribute to the design process and help eliminate some of the more tenuous aspects of model description. In addition, the graphs provide a useful vehicle for improved communication and shared information among those working on a project.

By all accounts, the survey analysts and project managers who work on the surveys cited above found the contributions described in this paper extremely valuable. It will be through continued and expanded use that additional benefits and applications will arise.

6. Pierce and Gillis Paper

This paper is a superb case study for the development of effective edit cells and related tolerances. This report can be a textbook study. One of the author's primary objectives is stated clearly at the onset: "A major task in setting edit tolerances is to ensure sensitivity without generating unnecessarily large quantities of 'false positive' exceptions." They have an excellent test environment because there is an unequivocal response question and the "truth" can always be determined (by subsequent revision) so the appropriateness of the edit can be evaluated.

Note that the authors clearly have a quintessential micro-editing requirement. Namely, their intended product is a longitudinal file of individual records of bank deposits.

After describing the underlying survey environment, the authors described their step-by-step process to design effective edit tolerances. They described how they developed the definition of edit cell and how they had to combine cells to get the proper break between cell size and homogeneity. They next described the model to predict (forecast) reported deposits, discussed alternative models and provided cogent reasons for each decision along the way. Following that, the authors describe their procedure for setting cell edit tolerances. After details of the edit system were decided upon, they were able to test various options based on the 1991-92 edit experience.

They took a major step in couching their analysis in terms of Type I and Type II errors to evaluate findings. The authors provided extensive tables and descriptions of their analysis. It is interesting to note that the current system has too many Type II errors, and future work will introduce refinements to achieve a lower rate.

Although one rarely comes across such a well-suited environment to test edit procedures and evaluate performance, this report is valuable in describing how to proceed under ideal circumstances. Using this ideal as a guide, one can modify procedures and change directions based on information actually available when attempting to apply the methods described in this report to other surveys.

7. Concluding Remarks

Both of these papers have a great deal to offer the reader. The first clearly illustrates how graphics can be applied to actual editing issues. One would hope that the examples here can suggest methods which can be applied to other surveys. The second paper is an excellent case-study for developing edit tolerances and evaluating them. This paper also can serve as a guide in helping others plan their own evaluation projects.